Reliability: An Essential Quality of a Research Tool

Juliana Linnette D'Sa

In research studies, just as in the case of me chanical and electronic equipment, accuracy in measurement is most important. Therefore, it is the aim of every researcher to minimise errors in measurement to the maximum extent. One way of achieving accuracy is by ensuring that the measuring tool or instrument used such as schedules, questionnaire, opinionnaire, etc. is reliable. Reliability is 'the degree of consistency with which an instrument measures the attribute it is supposed to be measuring (Polit & Hungler, 1995). Reliability is also described as the 'dependability' or 'trustworthiness' of a measuring instrument. It is a statistical concept and is expressed by means of reliability coefficient 'r' which can range from -1.00 through + 1.00; the higher the coefficient, the more the consistency between the measures.

But what should be the right correlation coefficient? Ideally the correlation coefficient computed between the two sets of scores should be dose to + 1.00 as it represents a perfect correlation. The correlation coefficient can range from -1.00 (a perfect negative relationship) through zero to +1.00 (a perfect positive correlation). What correlation coefficient is acceptable for an instrument to be reliable is debatable. Most researchers use the standards of Nunally (1978) according to which a newly developed instrument should have a reliability coefficient of at least 0.70 to be considered acceptable. If the instrument is an established one, it should measure 0.80 or more.

Reliability Measures

There are three different types of reliability measures. Stability (test-retest) internal Consistency and Equivalence. There are also some techniques and formulas used for determining reliability such as split-half technique, Kuder Richardson formula, Cronbach alpha.

Stability is the consistency with which the same results are obtained on repeated administrations of the same instrument. To determine the stability of the measure, the researcher uses the test-retest method by administering the test to the same group of individuals over a period of time. While administering the two tests the researcher keeps the conditions of testing, like the time of the day, room comfort and lighting constant in order to minimise the error rate that can affect the reliability scores. Comparison of the two sets of scores so obtained is done by computing the coefficient *r.* The higher the reliability coefficient, the more stable the measure is.

Notte et al (2012) determined the test - retest reliability of the Patient Perception of Urgency Scale, a patient-reported outcome instrument intended to measure the intensity of urgency associated with each urinary or incontinence episode. The test -retest reliability was high based on intra-class correlation of 0.95. Boltz et al (2009) established the reliability of the Geriatric Institutional Assessment Profile (GIAP), a self-administered survey of hospital readiness to geriatric programmes. Findings demonstrated the intra class correlation coefficients of the GIAP scales ranged between 0.82 and 0.92. D'Sa (2007) also used this method to establish the reliability of the Four Marker stations of the Objective Structured Clinical Examination on antenatal care, which yielded a reliability coefficient of 0.82.

There are, however, drawbacks in this method. Firstly, the individual's response-on the second administration of the test may be influenced by the memory of the responses in the first administration. This is likely to yield a higher correlation coefficient. Secondly the time lag between the two testing events is important; it should be neither too long nor too short. Waltz, Strickland & Lenz (1991) recommend a two-week interval between most testing. But the researcher is in a better position to decide the ideal length of time, based on the knowledge of the environment and the subjects. Thirdly, the individuals may find the procedure boring on the second administration and may respond haphazardly, thereby resulting in a low correlation coefficient.

Internal consistency is another method of determining reliability. In this, all the items on an instrument are made to measure only the critical attribute. One of the methods used is known as split-half technique and is popular with the researchers as the instrument is administered only once and therefore economical. More importantly, it assesses the source of measurement error i.e. sampling of items.

The author is Professor, Research in Nursing, Yenepoya University, Deralakatte, Mangalore.

Sampling errors are one of the main factors that influence reliability.

In *split-half* technique, the test is split into two halves (two sets). Although different methods are used for splitting the test into two halves like the odd-even method and the first half and the second half method, the most acceptable method used is the odd-even method. In this method, all the odd-numbered items form one set while the even-numbered items form another set. The scores of the two half tests are then used to compute a correlation-coefficient. Take for example a 30- item test. When this test is split into two halves, the odd items (1,3,5,7,..) form one set, and the even items (2,4,6,8..) form another set. Each individual answering the test thus receives two scores. The number of correct answers on all the odd numbered items constitutes one score and the number of correct answers on all even numbered items constitute another score.

The correlation coefficient thus computed yields results only for 15 items, in this case (half test), as the instrument is divided into two halves. To adjust the correlation coefficient for the entire test the Spearman-Brown Prophecy Formula is used.

It should be noted that in this method the tests should be long. The longer the scales the more reliable they are, other things being equal. Therefore, it is essential that the scale must be of sufficient length.

One of the disadvantages of split-half method is that different reliability estimates can be obtained by using different 'splits' like odd-even halves and the first-half, second-half. D'Sa (2002) reported use of split-half technique and Spearman Brown Prophecy Formula to establish the reliability of a clinical reasoning ability questionnaire on pregnancy- induced hypertension. The reliability coefficient was 0.87. The tool was considered reliable.

Although split-half technique is commonly used, some researchers use other methods like the Kuder-Richardson and the coefficient alpha (Cronbach's alpha). Another method of estimating internal consistency reliability is devised by Kuder and Richardson (1937). For using the Kuder- Richardson (KR) method, it is essential that all items should be homogenous and that all correct answers should be scored as +1 and all incorrect answers as zero. The KR-20 is the basic formula and KR-21 is the revised formula. The KR-21 is most widely used as it does not demand an item analysis worksheet, as does the KR- 20, which requires a knowledge of difficulty value for each item (proportion of correct answers).The KR-21 requires only the mean of the total test scores, standard deviation of the total score and the number of items in the test.

Huizinga et al (2008) developed and validated a Diabetes Numeracy Test (DNT), a scale to specifically measure numeracy skills used in diabetes. This 43-item DNT was highly reliable as determined by internal consistency (KR-20= 0.95). KR-20 is better than KR-21, the main reason being that KR-21 assumes that each item has "same" difficulty level.

Coefficient Alpha: Cronbach 1951) developed a Coefficient alpha to assess the internal consistency of an instrument by correlating each item with all other possible combination of items and thereby determine whether the items are homogeneous, that is, related to one concept. If the items are heterogeneous, they need to be placed into sub-scales through factor analysis or through other means. The alphas of the sub-scales are then determined.

Coefficient alpha is used when items have different weightages, for example, a five-point Likert-scale where scores range from 0-4 or 1-5. A five-point Likert scale opinionnaire on dual role of nurses, Balachandran (2008) yielded an alpha 0.90, which indicated a high reliability. D'Sa (2007) developed a 35 items Motivation for Learning Scale, to 'determine the motivation of baccalaureate nursing students, which was a forced-choice type, with three alternatives. Each of the alternatives had a score that ranged from 1-3. The alpha for the whole test was 0.809. Several other authors of tools have used the coefficient alpha.

Equivalence:

The measure of Equivalence is estimated in two different ways. One by using parallel forms /alternate forms of tests and determining the correlation coefficient and the other by agreement between measurements.

The two parallel / alternative forms of a single test is administered randomly in immediate succession to the same individuals. The order in which the two tests are administered should vary to balance the error rate for fatigue or testing effects. If the tests are administered in different orders to different individuals, the error rate is assumed to be equally distributed, which will then not affect the correlation coefficient computed between the two sets of scores which would yield the measure of reliability, known as coefficient of equivalence. In this situation, the demerits of the test-retest method, that is, the practice effect and the memory effect are automatically controlled because the second form of the test is similar to the first form, but not

identical.

It is often difficult to make both forms equivalent. Freeman (1962) has listed the following criteria for judging whether the two forms of the test are parallel: In both forms:

- the number of items should be same
- the items should have uniformity regarding the content, the range of difficulty and adequacy of sampling
- the index of difficulty of items should be similar
- the items should be of equal degree of homogeneity (shown through item-item correlation or by correlating each item with subset scores or with total test scores).
- the mean and standard deviation should be equal or near equal,
- the mode of administration and scoring should be uniform.

The other methods of determining equivalence is through inter-rater (inter-observer) reliability or intra-rater reliability. Inter-rater reliability is done by having two or more trained observers observing the same phenomenon at the same time and independently recording the relevant variables using one instrument (like a rating scale, or checklist of a procedure). In intra-rater reliability, one rater uses the same instrument for observation twice. The scores are then computed and an index of equivalence is determined through a correlation of coefficient or a function of agreement.

The percentage of agreement is determined using the formula:

$$\frac{\text{Number of agreements}}{\text{Number of agreements + Number of disagreements}}$$

Brennan & Hays (1992) suggest the use of Kappa statistics for interrater reliability. When three or more raters rate independently, Cronbach's coefficient alpha can be used (Waltz, Strickland & Lenz, 1991).

Factors Influencing Reliability

The reliability of scores of an instrument is influenced by a large number of factors, the extrinsic and the intrinsic (Singh,1986). Variability in the group and environmental conditions are some examples of extrinsic factors. The length of the instrument, scorer reliability and homogeneity of the items are some of the examples of intrinsic factors.

Extrinsic factors which are described below are those factors that lie outside the instrument.

Group Variability: A group that has a range of abilities, or are heterogeneous in abilities, can give a high reliability. On the other hand, if the group is homogenous in ability, the reliability may be low.

Guessing by the individual: Using the instrument by guess work tend to give high total scores and yield a spuriously high reliability. As guessing can yield different scores for different individuals it may also cause measurement error. For example, by guessing one may get 80 percent of the answers correct, while another may get only 40 percent of the answers correct. Such differences in the obtained scores may cause error score, thereby lowering the reliability of the scores.

Environmental conditions: Environmental conditions like lighting, noise levels, temperature, relative level of the testing environments should remain the same. If there is variation it can lower the reliability of the instrument.

The main intrinsic factors that affect the reliability of the instrument are:

Spread of scores: Reliability coefficient is directly influenced by the spread of soeres in the group being tested. The larger the spread of scores, the higher the estimate of reliability, other things being equal.

The length of the test/ instrument: Longer instruments yield higher reliability coefficient, while shorter instruments yield lower reliability coefficient. This is because longer instruments will provide a more adequate sample of the behaviour being measured. A longer test also tends to lessen the influence of chance factors such as guessing, which cause measurement errors.

Homogeneity of items: Homogeneity includes item reliability (item-item correlation) and the homogeneity of function or trait that is being measured. When all the items measure the same trait, inter-item correlation is higher, and the reliability of the test is high, and vice versa.

Difficulty Value of items: The items having indices of difficulty of 0.5 or close to it yield higher reliability. If the items are too easy or too difficult, it can affect reliability.

Discrimination Value of items: When the instrument is composed of items that discriminate, the item-total test, correlation is likely to be high. Therefore, the reliability is also likely to be high and vice versa.

Scores Reliability: It means how closely two or more

scorers agree in scoring or rating the same set of responses. If the agreement is low, the reliability is likely to be low.

How Reliability can be improved: There are ways in which reliability of an instrument can be improved. These include:

- Having a heterogeneous group of subjects with wide variability in the ability or trait being measured.
- Keeping items in the test homogenous.
- Keeping the length of the test appropriate.
- Choosing items with moderate difficulty and discriminating indices.
- Controlling the environmental factors

Conclusion

Reliability is an essential quality of an instrument. Reliability coefficients are determined by several methods, like stability, internal consistency and equivalence. It is essential that the researcher uses the appropriate method when estimating the reliability of a tool. It is therefore important to consider the factors that affect reliability before developing a research tool.

References

1. Balachandran R. Dual role of nurses as educator and practitioner. *The Nursing Journal of India* 2008; 99(10):231 -233.
2. Boltz M, Capezuti E, Kim H Fairchild S. Test- retest reliability of the geriatric institutional assessment profile. *Clinical Nursing Research* 2009; 18(3): 242-52
3. Bremen PF, Hays BJ. The Kappa statistic for establishing interrater reliability in the secondary analysis of qualitative clinical data. *Research in Nursing and Health* 1992; 15:153-58
4. D'Sa JL. Evaluation of a problem-based learning package on pregnancy induced hypertension for BSc Nursing students. *The Nursing Journal of India* 2002, 93(11): 261-62
5. D'Sa JL. Effectiveness of a problem-based learning package on clinical reasoning, clinical skills and attitude of students towards care of antenatal clients in selected institutions of Kamataka state, Doctoral thesis, 2007. Manipal University, Manipal
6. Freeman FS. Theory and Practice of Psychological Testing. 1962. New York: Holt, Rinehart & Winston (Indian edition)
7. Huizinga MM, Elasy TA, Waliston KA, Cavanaugh K, Davis D, Gregory, R P, Fuchs LS, Malone R, Cherrington A, DeWalt DA, Buse J, Pignone M, Rothman RL. Development and validation of the Diabetes Numeracy Test (DNT) BMC Health Services Research 2008, 8:96
8. Kuder GF, Richardson MW. The theory and estimation of test reliability. *Psychometrika* 1937; 2:151-60
9. Singh AK. Tests Measurements and Research Methods in Behavioural Sciences. 1986; Delhi; Tata McGraw-Hill Publishing Company Limited
10. Polit DF, Hungler BP. Nursing Research Principles and Methods. 5th edn, 1995. Philadelphia; JP Lippincot
11. Noite SM, Marshall TS, Lee M, Hakimi Z, Odeyemi I, Chen WH, Revicki D A. Content validity and test-retest reliability of patient perception of intensity of urgency scale (PPIUS) for overactive bladder BMC Urology 2012; 12:26
12. Nunally J. Psychometric Theory. 1 978. New York: McGraw-Hill
13. Waltz CF, Strickland OL, Lenz ER. Measurements in Nursing Research. 2nd edn, 1991. Philadelphia: FA Davis

Lodging at TNAI Headquarters Made Easier

TNAI Hqrs has expanded its capacity to accommodate more of TNAI members visiting Delhi. The TNAI members including students visiting Delhi on official or professional tours can avail the lodging facility, within the TNAI Hqrs premises at reasonable charges. The per day charges are as under:

TNAI Members: **Rs. 400/-**	SNA Members: **Rs. 150/-**	Non-Members: **Rs. 700/-**
Children below 5 yrs: No charges	Children 6-12 yrs: **Rs. 100/-**	

However, due to limited beds, interested members may get the booking done in advance.

Mrs Sheila Seda, Secretary-General, TNAI

Attention Advertisers !

Advertisers of the Admission Notices in TNAI Bulletin for the academic year 2013-2014 for Schools / Colleges of Nursing are required to submit the copy of Indian Nursing Council (INC) recognition certificate along with the advertisement and payment, otherwise the advertisement shall be summarily rejected.

- Chief Editor